УДК 681.3, 621.3

O.M. YATSKO, YU.O. USHENKO, O.V. OLAR

REVIEW OF INTELLIGENT DATA ANALYSIS TO WEB-DEVELOPMENT APPLICATIONS

Yuri Fedkovich Chernivtsi National University, Chernivtsi,
Kotsyubinsky Street 2, Chernivtsi, Ukraine, e-mail: o.yacko@chnu.edu.ua

Abstract. The essence of intelligent data analysis (data mining), methods of intelligent data analysis are considered. The field of application of intelligent data analysis and existing systems are analyzed. Conclusions were made regarding the prospects of using methods of intelligent data analysis in web development.
Ключові слова: data mining; methods of intelligent data analysis are considered, business process, web development.

Анотація. Розглянуто сутність інтелектуального аналізу даних (data mining), методи інтелектуального аналізу даних. Проаналізовано сферу застосування інтелектуального аналізу даних та існуючі системи. Зроблено висновки стосовно перспектив використання методів інтелектуального аналізу даних у веб-розробці.
Keywords: data mining; methods of intelligent data analysis are considered business process, web development.

DOI: 10.31649/1681-7893-2022-43-1-36-42

INTRODUCTION

Modern information technologies have changed life and led to the restructuring of the structure of society. The current stage of the development of society is a change from a mass consumption society to an information society, which has the following characteristics:

- the dynamics of life have increased, that is, the time interval between the appearance of new data and their assimilation by people has decreased;
- data transformation systems are becoming the most important structuring factor in society, covering almost all spheres of human activity;
- there is a relative decrease in the need for raw materials and energy due to an increase in the speed and volume of information flows;
- at this point, the primary factor of production is knowledge, not capital;
- key technologies are organizational, operational, informational, not machine-based;
- a person actively creates information, not just passively perceives it;
- the definition of the category «knowledge» is changing – there is a transition from possessing the knowledge to managing knowledge.

In connection with the improvement of data transmission and storage technologies, colossal flows of information have been obtained in various spheres of human activity. At present, without productive processing, streams of raw data are of no use to anyone. The specifics of modern requirements for such processing are as follows:

- data has unlimited volume;
- data are heterogeneous (quantitative, qualitative, textual);
- results must be specific and clear;
- raw data processing tools should be easy to use.

1. THE CONCEPT OF INTELLIGENT DATA ANALYSIS

In connection with the growth of the flow of data in modern times, the problem of processing this flow and identifying regularities arises. These regularities play an important role in evaluating the strategy and tactics of this or that institution, as well as identifying its potential.

A person cannot analyze extremely large volumes of data on his own, but the need for such an analysis is quite obvious because in these data flows there is the knowledge that can be used during decision-making. Mathematical statistics for a long time claimed the role of the main tool of data analysis but did not correspond to the problems that arose. In this connection, there was a need to develop new modern methodologies for data flow processing and analysis. Intelligent data analysis (IAD) became such a new methodology. The advantages of intelligent data analysis were:

- rapid accumulation of extremely large data;
- general computerization of business processes;
- penetration of the global computer network into all spheres of activity;
- progress in the field of information technologies: improvement of databases and data warehouses;
- progress in the field of production technologies: the rapid growth of computer productivity, volumes of drives, implementation of Grid systems.

Algorithms used in IAD require a large number of calculations. Previously, this was a deterrent to the widespread practical use of IAD, but today's increase in the performance of modern processors has removed the severity of this problem. Now, in an acceptable time, it is possible to perform a qualitative analysis of hundreds of thousands and millions of records. IAD is an interdisciplinary field that arose and developed based on such sciences as applied statistics, pattern recognition, artificial intelligence, database theory, etc. [1, p. 54].

Intelligent data analysis (IAD) or data mining (discovery-driven data mining) is the process of discovering in primary data previously unknown, accessible, practically useful, and non-trivial interpretations of knowledge necessary for decision-making in various spheres of human activity. Visual IAD tools allow data analysis by subject specialists who do not have appropriate mathematical knowledge [2, p. 382].

2. METHODS OF INTELLIGENT DATA ANALYSIS

Data mining methods are divided into [3, p. 390]:

- statistical: descriptive analysis, correlation and regression analysis, factor analysis, dispersion analysis, component analysis, discriminant analysis, time series analysis;
- cybernetic: artificial neural networks, evolutionary programming, genetic algorithms, associative memory, fuzzy logic, decision trees, expert knowledge processing systems.

Modern IAD technology is based on the concept of templates that reflect fragments of multifaceted relationships in data. These patterns are patterns inherent in subsamples of data that can be compactly expressed in an understandable form. The search for patterns is carried out by methods that are not limited by the framework of a priori assumptions about the structure of the sample and the types of distributions of the values of the analyzed indicators [2, p. 94].

One of the principles of intelligent data analysis is the non-triviality of search patterns. This means that the found patterns should reflect non-obvious, unexpected regularities in the data, components of the so-called hidden knowledge. Raw data (raw data) contain knowledge, which, with the proper interpretation, can reveal important information.

At the beginning of its development, the use of neural networks in data analysis caused mixed opinions due to their shortcomings, such as the complexity of the structure, too long a training period, and poor interpretability.

But they were justified by a set of positive qualities, such as a low error rate, constant improvement, and optimization of various algorithms for learning networks, an algorithm for obtaining rules, and an algorithm for simplifying networks, which make neural networks an extremely promising direction in the field of data analysis [2, p. 381].

The areas of use of neural networks are forecasting, classification, clustering, adaptive management, creation of expert systems, automation of image recognition processes, processing of analog and digital signals, synthesis and identification of electronic circuits and systems, etc. [4].

Clustering, or natural classification, is the process of grouping objects with similar characteristics into groups. In contrast to the usual classification, where the number of groups of objects is fixed, here neither the groups nor their number are determined in advance and are formed during the operation of the system, based on the proximity of the objects [5, p. 579].

Clustering is used to solve many applied problems – from image segmentation to economic forecasting and combating electronic fraud.

The task of clustering is relevant since the growing accumulation of data leads to the need for their classification. During the analysis of objects or phenomena, it becomes necessary to take into account an increasing number of parameters, therefore there is a task of developing and applying methods that specialize in the classification of multidimensional data [5, p. 579].

Often there is a need to somehow classify the data or find patterns in it. This can be achieved using both clustering algorithms and neural network methods, as well as fuzzy network processing methods [4, 6, 7].

Neural network models can be conventionally divided into three types, such as:

1) direct propagation networks are one of the most common architectures used in forecasting and pattern recognition;

2) networks with feedback, which is used to optimize calculations and associative memory;

3) self-organizing networks containing models of adaptive resonance theory and Kohonen models and used for cluster analysis [5, p. 579].

Currently, active development of clustering algorithms, which can process very large databases, is underway. Algorithms, where hierarchical clustering methods are integrated with other methods, have been developed. The most relevant algorithms include BIRCH, CURE, ROCK, Chameleon, and Kohonen [8, p. 72].

There are many types of data analysis based exclusively on neural networks, but two of them are the most popular. They are based on self-organizing neural networks and fuzzy networks.

1. Data analysis based on a self-organizing neural network. The self-organizational process is the process of learning without a teacher.

During such learning, the training set consists of the values of the input variables, and in the learning process, there is no comparison of the neuron outputs with the desired values. We can say that such a network learns to understand the data structure [9, p. 231].

The idea of the Kohonen network belongs to the Finnish scientist Toivo Kohonen. The principle of operation of these networks consists in introducing into the learning rule of neuron information about its location, that is, maps of the location of neurons are made.

Kohonen self-organizing maps are used for modeling, forecasting, finding patterns in large data sets, identifying sets of independent features, and compressing information.

2. Data analysis (data mining), based on a fuzzy neural network. Fuzzy neural networks are based on the idea of using an existing sample of data to determine the parameters of membership functions, conclusions are formulated based on a fuzzy logic apparatus, and learning algorithms of neural networks are used to find the parameters of membership functions. Such systems can use previously known information, learn, acquire new knowledge, forecast time series, and perform image classification. But one of the main advantages is the visibility of the operation of such a network for the user.

Each of the considered types of neural networks has its advantages and disadvantages for intelligent data analysis, so it is appropriate to compare the Kohonen neural network in the group of types of intelligent data analysis based on neural networks.

The main difference between Kohonen networks and other types of neural networks lies in clarity and ease of use. These networks allow for the simplification of a multidimensional structure and can be considered as one method of projecting a multidimensional space into a lower-dimensional space. Another fundamental difference between Kohonen networks is that while all other networks are designed for supervised learning tasks, the Kohonen network is primarily designed for unsupervised learning, meaning that the network learns to understand the structure of the data itself.

One of the most significant disadvantages of Kohonen's neural network is that the corresponding algorithm does not provide for determining the number of clusters. But it can function in conditions of interference because the number of clusters is fixed in advance.

So, Kohonen's self-organizing neural network can be one of the foundations of an adequate algorithm compared to other types of neural networks designed for cluster data analysis and competes with modern algorithms, but none of the existing pure models meet modern requirements.

3. IAD SOFTWARE

Subject-oriented analytical systems

The broadest subclass of data systems that have become widespread in the field of financial market research is called "technical analysis." This is a collection of several dozen methods of forecasting price dynamics and choosing the optimal structure of an investment portfolio, based on various empirical models of market dynamics. These methods usually use a simple statistical apparatus, but take into account the established specifics of the industry as much as possible (professional language, systems of various indexes, etc.).

Statistical packages

The latest versions of almost all known statistical packages include, along with traditional statistical methods, also elements of intelligent data analysis, but the main attention is paid to classical methods - correlation, regression, factor analysis, etc. The disadvantage of these systems is that the user must have special training. Another disadvantage of statistical packages is that it limits their use in intelligent data analysis.

Examples of the most powerful and widespread statistical packages are SAS (SAS Institute company), SPSS (SPSS), STATGRAPICS (Manugistics), STATISTICA, STADIA, etc.

Neural networks

This is a large class of systems, the architecture of which has an analogy with the construction of nervous tissue from neurons. In one of the most common architectures, the work of neurons is simulated as part of a hierarchical network, where each neuron of a higher level is connected by its inputs to the outputs of neurons of a lower layer. The neurons of the lowest layer are fed the values of the input parameters, based on which it is necessary to make decisions, predict the development of the situation, etc. These values are considered signals transmitted to the next layer, weakening or strengthening, depending on the numerical values (weights) attributed to the interneuron connections. As a result, some value is produced at the output of the neuron of the highest upper layer, which is considered a response - the reaction of the entire network to the entered values of the input parameters.

The main disadvantage of the neural network paradigm is the need for a very large training sample.

Examples of neural network systems are BrainMaker (CSS), NeuroShell (Ward Systems Group), and OWL (HyperLogic). Their cost is quite high.

Reasoning systems based on similar cases

The idea of case-based reasoning (CBR) systems is extremely simple at first glance. To make a forecast for the future or choose the right decision, these systems find close analogs of the current situation in the past and choose the same answer that was right for them. Therefore, this method is also called the "nearest neighbor" method. Recently, the term memory-based reasoning has also become popular, which emphasizes the fact that a decision is made based on all the information stored in memory.

Their main drawback is that they do not create any models or rules that generalize previous experience at all - in choosing a solution, they are based on the entire array of available historical data, so it is impossible to say based on which specific factors CBR systems build their responses.

Examples of systems using CBR are KATE tools (Acknosoft, France), and Pattern Recognition Workbench (Unica, USA).

Decision trees

Decision trees are one of the most popular approaches to solving problems of intelligent data analysis. They create a hierarchical structure of rules of the type «IF... THEN...» (if-then), which has the appearance of a tree.

Decision trees are fundamentally incapable of finding the «best» (most complete and most accurate) rules in the data. They implement the naive principle of sequential review of signs and actually «touch» parts of real patterns, creating only the illusion of a logical conclusion.

The most famous systems of this method are See5/Z5.0 (RuleQuest, Australia), Clementine (Integral Solutions, Great Britain), SPINA (University of Lyon, France), IDIS (Information Discovery, USA), KnowledgeSeeker (ANGOSS, Canada).

Genetic algorithms

Data mining is not the main area of application of genetic algorithms. It is a powerful tool for solving various combinatorial and optimization problems. However, genetic algorithms have become part of the standard toolkit of intelligent data analysis methods.

The first step in the construction of genetic algorithms is to encode the original logical patterns in a database called chromosomes, and the entire set of such patterns is called a population of chromosomes. Next, to implement the concept of selection, a method of comparing different chromosomes is introduced. The population is processed using the procedures of reproduction, variability (mutations), and genetic composition. These procedures mimic biological processes. The most important among them are random mutations of data in individual chromosomes, transitions (crossing over) and recombination of genetic material contained in individual parental chromosomes, and gene migration.

In the course of the procedures at each stage of evolution, populations with more and more perfect individuals emerge.

Genetic algorithms are convenient because they are easy to parallelize. For example, you can divide the generation into several groups and work with each of them independently, exchanging several chromosomes from time to time. There are also other methods of parallelizing genetic algorithms.

The criteria for selecting chromosomes and the procedures used are heuristics and by no means guaranteed to find the "best" solution.

An example can be the GeneHunter system of Ward Systems Group.

Algorithms of limited enumeration

Algorithms of limited search of the past were proposed in the mid-60s of the 20th century. M.M. Bongard for finding logical patterns in data. Since then, they have demonstrated their effectiveness in solving a multitude of problems from various fields.

These algorithms calculate the frequencies of combinations of simple logical events in subsets of data.

The most striking modern representative of this approach is the WizWhy system from WizSoft.

Systems for visualization of multidimensional data

In such systems, the main attention is focused on the tolerance of the user interface, which makes it possible to associate various parameters of the scatter diagram of objects (records) of the database with the analyzed indicators. Such parameters include color, shape, orientation relative to its axis, dimensions, and other properties of graphic elements of the image. In addition, data visualization systems are marked with convenient tools for scaling and rotating images.

4. APPLICATION OF METHODS OF INTELLIGENT DATA ANALYSIS

The field of application of methods of intelligent data analysis is not limited by anything, but these methods are primarily used today by commercial enterprises that connect their activities with data warehouses (Data Warehousing). Entrepreneurs have realized that with the help of methods of intelligent data analysis, they can get tangible advantages in the competition. The experience of many enterprises shows that the profit from the use of intelligent data analysis can reach more than 100%.

Fields of application of intelligent data analysis:

Retail

Retail businesses today collect information on each purchase using store-branded credit cards and an automated tracking system. Typical tasks that can be performed with data mining in the retail industry are:

- shopping cart analysis (similarity analysis) is designed to identify products that customers tend to buy together. Knowledge of the shopping cart is necessary for improving advertising, developing a strategy for creating stocks of goods and methods of their location in trading rooms;
- the study of time patterns helps trading enterprises to make decisions about the creation of commodity stocks;
- the creation of predictive models enables trade enterprises to learn about the nature of the needs of different categories of customers with certain behavior, for example, buying goods from famous designers or visiting sales. This knowledge is needed for the development of precisely targeted, economical measures for the promotion of goods on the market.

Banking

Advances in the technology of intelligent data analysis are used in banking to perform the following tasks:

- credit card fraud detection. Analyzing past transactions that later turned out to be fraudulent, the bank reveals some stereotypes of such fraud;
- customer segmentation. By dividing customers into different categories, banks make their marketing policy more targeted and effective, offering different types of services to different groups of customers;
- predicting changes in clientele. Data mining helps banks build predictive models of the value of their customers and appropriately serve each category of customers.

Telecommunications

In telecommunications, data mining techniques help businesses advance their marketing and pricing programs to retain existing customers and attract new ones. Typical measures are:

- analysis of records on detailed characteristics of calls. The purpose of such an analysis is to identify categories of customers with similar stereotypes of using their services and develop attractive sets of prices and services;
- detection of customer loyalty. Intelligent data analysis can be used to determine the characteristics of customers who have used the services of this company at least once, with a high probability of remaining loyal to it. As a result, the funds allocated for marketing can be spent on where the profit will be the greatest.

Insurance

Insurance companies have been accumulating large amounts of data for many years. In this field, methods of intelligent data analysis:

- fraud detection. Insurance companies can reduce the level of fraud by looking for certain stereotypes in insurance claims that characterize the relationship between lawyers, doctors, and applicants;
- risk analysis. By identifying combinations of factors associated with paid claims, insurance companies can reduce their liability costs.

Medicine

There are many expert systems for making medical diagnoses. They are built mainly based on rules describing the combination of symptoms of diseases.

With the help of such rules, they recognize not only what the patient is sick with, but also how to treat him. The rules help to choose means of medicinal influence, determine indications – contraindications, navigate medical procedures, create conditions for the most effective treatment, predict the results of the prescribed course of treatment, etc. Technologies of intelligent data analysis make it possible to detect these patterns, which are components of the specified rules.

Molecular genetics and genetic engineering

Perhaps the most urgent task is to identify regularities in experimental data in molecular genetics and genetic engineering. Here it is formed as a definition of markers, which are understood as genetic codes that control certain phenotypic characteristics of a living organism. Such codes can contain hundreds, thousands, or more related elements.

Large funds are allocated for the development of genetic research. Recently, there has been a special interest in the application of intelligent data analysis methods in this field.

Applied Chemistry

Methods of intelligent data analysis are widely used in applied chemistry (organic and inorganic). Here, the question of clarifying the specifics of the chemical structure of certain compounds, and their defining properties, often arises. Such a task is especially relevant when analyzing complex chemical compounds, the description of which includes hundreds and thousands of structural elements and their connections.

Therefore, the potential of intelligent data analysis gives impetus to the expansion of the limits of the application of this technology in the modern world of computer technologies. Regarding the prospects of intelligent data analysis, the following directions of development are possible:

- selection of types of subject fields with their corresponding heuristics, the formalization of which will facilitate the solution of relevant problems of intellectual analysis of data belonging to these fields;
- creation of formal languages and logical means, with the help of which reasoning will be formalized and automation of which will become a tool for solving problems of intellectual data analysis in specific subject areas;
- creation of methods of intellectual data analysis capable not only of extracting regularities from the data, but also of forming some theories based on empirical data;
- overcoming the significant lag behind the possibilities of instrumental means of intellectual data analysis from theoretical achievements in this area.

The main feature of intelligent data analysis is the combination of a wide range of mathematical tools (from classical statistical analysis to new cybernetic methods) and the latest achievements in the field of information technology. In the technology of intelligent data analysis, formalized methods and methods of informal analysis are harmoniously combined, that is, quantitative and qualitative data analysis.

CONCLUSIONS

Intelligent data analysis systems are used as a mass product for business applications and as tools for conducting unique research (genetics, chemistry, medicine, etc.).

Leaders in intelligent data analysis associate the future of these systems with their use as intelligent applications embedded in corporate data warehouses.

REFERENCES

1. Xianjun Ni Research of Data Mining Based on Neural Networks // World Academy of Science, Engineering, and Technology. – 2008. – №39. – P. 381-384.
2. K. Ilyashenko. Information methods of intellectual data analysis/ K. Ilyashenko // Economic analysis: coll. of the science works / Ternopil National University of Economics: TNEU Publishing House «Economic Thought». – 2010 – №7. – P. 390-393.

3. Xu R. and Wunsch D. II. Survey of Clustering Algorithms. IEEE TRANSACTIONS ON NEURAL NETWORKS, – VOL. 16. – №3, 2005, – PP. 645-678.
4. Leshchynsky O.L., Ishchenko A.O. The use of neural networks in the process of intellectual (cluster) data analysis / O.L. Leshchynskyi, A.O. Ishchenko // Economy and society. - 2017 - No. 11. [Electronic resource]. – Access mode: http://www.economyandsociety.in.ua/journal-11/18-stati-11/1269-leshchinskij-ol-ishchenko-ao
5. Mandel I.D. Cluster analysis. – M.: Finances and Statistics, 1988. – 176 p.
6. Starykov A. Practical application of neural networks for classification (clustering) tasks, [Electronic resource]. – Access mode: http://www.basegroup.ru/neural/practice.htm, January 2000.
7. George Karypis. Chameleon: Hierarchical Clustering Using Dynamic Modeling / George Karypis, Eui-Hong (Sam) Han, Vipin Kumar // Computer. – 1999. – Vol. 32, №8. – P. 68-75.
8. Shumeiko A.A., Sotnyk S.L. Intellectual data analysis / A.A. Shumeiko, S.L. Centurion – Dnipropetrovsk: Belaya, 2012. – 212 p.
9. Olexander N. Romanyuk, Sergii V. Pavlov, and etc. "A function-based approach to real-time visualization using graphics processing units", Proc. SPIE 11581, Photonics Applications in Astronomy, Communications, Industry, and High Energy Physics Experiments 2020, 115810E (14 October 2020); https://doi.org/10.1117/12.2580212.
10. Leonid I. Timchenko, Natalia I. Kokriatskaia, Sergii V. Pavlov, and etc. \ "Q-processors for real-time image processing", Proc. SPIE 11581, Photonics Applications in Astronomy, Communications, Industry, and High Energy Physics Experiments 2020, 115810F (14 October 2020); https://doi.org/10.1117/12.2580230
11. Information Technology in Medical Diagnostics //Waldemar Wójcik, Andrzej Smolarz, July 11, 2017 by CRC Press - 210 Pages.
12. W. Wójcik, S. Pavlov, M. Kalimoldayev. Information Technology in Medical Diagnostics II. London: (2019). Taylor & Francis Group, CRC Press, Balkema book. – 336 Pages.
13. Intellectual Technologies in Medical Diagnosis, Treatment and Rehabilitation: monograph / [S. In Pavlov, O.G. Avrunin, S.M. Zlepko, E.V. Bodyanskyi, etc.]; edited by S. Pavlov, O. Avrunin. - Vinnytsia: PP "TD "Edelveiss and K", 2019. -260 p. ISBN 978-617-7237-59-3
14. Intelligent Technologies of Computer Planning and Modeling in Medical Diagnosis, Treatment and Rehabilitation: monograph // edited by S.V. Pavlov, O.G. Avrunin, O.V. Hrushko - Zhytomyr: "Euro-Volyn" PE, 2021. - 202 p. ISBN 978-617-7992-15-7.

Надійшла до редакції 15.04.2022р.

YATSKO OKSANA – Ph.D., assistant professor of Computer Science Department, Yuriy Fedkovich Chernivtsi National University, Chernivtsi, Ukraine, ***e-mail: o.yacko@chnu.edu.ua***
m.gorskiy@chnu.edu.ua

USHENKO YURIY -D.Sc., Professor of Computer Science Department, Yuriy Fedkovich Chernivtsi National University, Chernivtsi, Ukraine, ***e-mail: y.ushenko@chnu.edu.ua***

OLAR OLEKSANDR – Ph.D., assistant professor of Computer Science Department, Yuriy Fedkovich Chernivtsi National University, Chernivtsi, Ukraine, ***e-mail: o.v.olar@chnu.edu.ua***